Offline Surgical QA with Decomposed Retrieval and Synthesis for Resource-Constrained Settings

Kiran Bhattacharyya BHATTACHARYYAKIRAN12@GMAIL.COM *Intuitive Surgical, Inc., Atlanta, GA, USA*

Abstract

Digital access to critical medical knowledge in resource-limited settings is often hindered by a lack of internet connectivity and the computational demands of AI systems. This paper introduces the Surgical Information Assistant, a fully deployable, large language model (LLM) -driven multi-agent system designed to provide reliable surgical information in offline, resource-constrained environments. Our system is powered by a workflow that orchestrates question decomposition, information retrieval, grounded generation, and information synthesis to perform complex reasoning on consumer-grade hardware. Grounded in the Open Manual of Surgery for Resource-Limited Settings, we evaluated DeRetSyn on a new question-answer (QA) dataset of over 14,000 surgical question-answer pairs. We compare our system to other alternatives, perform ablation experiments on components of the agentic system, and interrogate sensitivity to retrieval parameters. The results show that our agentic orchestration enables a compact 3B Llama model to achieve 63% top-1 accuracy, significantly outperforming both a baseline GPT-4o (42.5%) and a larger 8B Llama model with conventional RAG (53%). We further test whether this performance enhancement from agentic orchestration for information retrieval generalizes to the PubMedQA dataset. Additionally, the entire system consumes <3.5 GB of RAM and generates responses within 8-15 seconds working on a consumer laptop. Our work serves as a practical blueprint for how agent-based systems can empower small, efficient models for medical domain information retrieval and synthesis, offering a tangible application of AI technology that could help advance health equity. We will release our dataset, code base, and prompts to foster further research in deployable and responsible clinical AI.

Keywords: clinical AI, information retrieval system, medical education, offline AI systems

Data and Code Availability The datasets used in this study are: 1) the PubMedQA dataset Jin et al. (2019) and 2) a dataset that we create and refer to as the *OpenManualOfSurgeryQA* dataset. The code and data is publicly available at this Github repo. Please note that this repo is under active development and there may be breaking changes.

Institutional Review Board (IRB) This research does not require IRB approval since it is not "human subjects research" as it does not include activities that involve interaction with individuals or access to identifiable private information.

1. Introduction

Lack of access to modern technology, educational resource constraints, and limited opportunities for training have been identified as key challenges for surgical practice in resource-limited settings Achanga et al. (2025). Large language models (LLMs) offer a powerful way to democratize access to clinical knowledge with potential to improve surgical training and education. However, in many real-world clinical environments, especially in resource-constrained settings, the reliance of most LLMs on large-scale models, stable internet access, or high-performance hardware (e.g. GPUs) presents a significant barrier to adoption. Furthermore, even in clinical settings with advanced technology access, there are significant privacy concerns about sharing patient data with LLM service providers Zhao and et al. (2024). This work directly addresses these practical limitations by developing a tool designed to further equity in clinical AI by targeting environments where opaque reasoning, retrieval failures, lack of connectivity, and lack of patient privacy can hinder effective care Ke et al. (2025); Ng et al. (2025); Shi et al. (2025).

Agent-based systems that orchestrate multiple components to solve complex problems offer a promising paradigm for improving the reliability and addressing limitations of LLMs. A common approach is Retrieval-

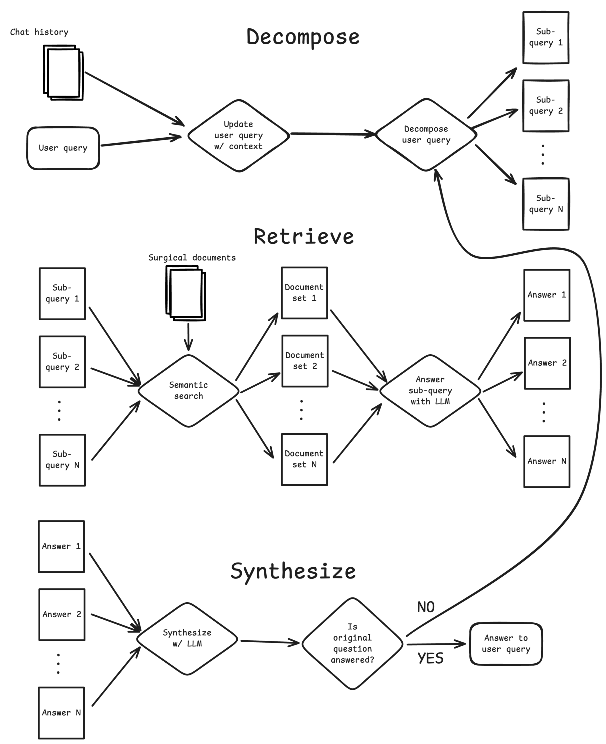

Figure 1: System design of the DeRetSyn pipeline.

Augmented Generation (RAG), which can be viewed as a simple two-stage workflow between a retriever and a generator agent to ground outputs in external knowledge Vrdoljak et al. (2025); Bunnell et al. (2025). More advanced systems now employ multi-agent, iterative workflows that decompose complex queries to emulate reasoning and improve interpretability Zhang and et al. (2024); Xu et al. (2024); Liu et al. (2025). However, these sophisticated agentic architectures often remain computationally intensive and reliant on cloud-based APIs, limiting their deployability where it is needed most.

To bridge this gap, we introduce the *Surgical Information Assistant*, an innovative and deployment-focused application designed for offline use. Built on the existing Open Manual of Surgery for Resource-Limited Settings Vanderbilt University Medical Center (2025), the system is designed to provide surgical information to user questions without an internet connection. It is powered by our multi-agent architecture, DeRetSyn (Decompose–Retrieve–Synthesize). This workflow orchestrates a collaboration between several specialized agents: 1) a Decomposition Agent that breaks down complex surgical queries into answerable sub-questions; 2) parallel Retrieval Agents that fetch relevant text from a local, FAISS-indexed corpus; and 3) a Synthesis Agent that generates and iteratively refines answers under the guidance of a Supervisor Agent which verifies completion, using a compact Llama-3.2-3B model.

In an evaluation on a new dataset of over 14,500 QA pairs, our DeRetSyn-powered 3B model achieves 63.0% accuracy, significantly outperforming both a baseline GPT-4o (42.5%) and a larger 8B model using standard RAG (53.8%). Furthermore, the agentic orchestration shows a consistent advantage also on the PubMedQA dataset Jin et al. (2019). The system we describe here is deployable on consumer-grade hardware without internet connectivity.

While DeRetSyn is not formally affiliated with the Open Manual of Surgery, it builds on its content under the Creative Commons license. We are making our entire framework publicly available, including all code, data, and prompts, to support further research and development in accessible clinical AI. Our experiments also explicitly measure deployment relevant factors such as latency, memory, and offline usability, and we discuss dataset biases, generalizability, and practical limitations to inform safe clinical translation.

Our contributions are:

- A deployable, open-source agent-based system for offline surgical QA, demonstrating an application of LLM and AI agents for further equity in clinical AI.
- Empirical evidence that a multi-agent workflow can significantly boost the performance of small, efficient language models, making them viable for clinical QA in resource-constrained domains.
- A QA dataset to support further research and improvements in deployable agent-based systems for clinical AI.

2. Methods

2.1. Knowledge Base for a Low-Resource Context

The foundation of the Surgical Information Assistant is a curated, offline knowledge base derived from the *Open Manual of Surgery for Resource-Limited Settings (OMSRS)*. We selected this corpus for two primary reasons: 1) its content is explicitly tailored for clinicians practicing without access to specialized equipment or resources, directly aligning with our goal of promoting health equity; and 2) its Creative Commons license permits redistribution, enabling the creation of a fully open-source and deployable system.

To prepare the corpus for retrieval, we parsed the source PDFs into text and segmented the content into chunks of 1000 characters with a 200-character overlap to maintain contextual continuity. These chunks were then converted into vector embeddings using the `all-MiniLM-L2-v2` model, chosen for its efficiency and strong performance in semantic search tasks. The resulting vectors were indexed into a local FAISS (Facebook AI Similarity Search) database, creating a fast and memory-efficient search index that can be deployed alongside the model on consumer-grade hardware and accessed via multiple threads. The total hard-disk space consumed by the index and the metadata was 60 MB.

2.2. The DeRetSyn Agentic Workflow

The core of the *Surgical Information Assistant* is DeRetSyn, a multi-agent workflow designed to emulate a structured reasoning process by breaking down complex problems into manageable steps (Figure 1). The system is coordinated by a lightweight, Python-based agentic controller that manages stepwise function calls for planning, retrieval, grounded generation, and synthesis. The workflow proceeds through the following specialized agents:

1. **Decomposition Agent**: An initial LLM-powered agent analyzes the user's query to identify and extract underlying sub-questions. For example, a query like "How do I manage a post-operative wound infection?" might be decomposed into "What are the signs of a wound infection?", "What are the common pathogens causing wound infections?", and "What are the first-line treatments for wound infection in a low-resource setting?".

2. **Parallel Retrieval Agents**: Each sub-question triggers an independent retrieval agent to perform a semantic search against the FAISS index. These agents run asynchronously to reduce overall response time, each retrieving the top-k most relevant text chunks for its assigned sub-question.

3. **Synthesis Agent**: For each sub-question and its retrieved context, a synthesis agent generates a concise, focused answer.

4. **Supervisor Agent**: A supervisor agent reviews the synthesized sub-answers to determine if the original query has been comprehensively addressed. If gaps are identified, this agent generates new sub-questions and repeats the retrieval-synthesis loop, accumulating knowledge over a maximum of three iterations. This iterative refinement allows the system to correct its course and construct a more complete answer.

5. **Final Synthesis**: Once the supervisor agent confirms that the query has been fully answered, a final LLM call synthesizes all verified sub-answers into a single, coherent response for the user. All intermediate sub-questions and answers are also provided to the user for review and transparency. If the system cannot fully answer the query, it explicitly states this and lists uncovered sub-questions.

Fallback Mechanism for Out-of-Domain Queries. To handle queries that fall outside the scope of the primary surgical corpus, the system includes an optional, fallback mechanism. If the DeRetSyn workflow fails to find a satisfactory answer after three iterations, it can query a pre-indexed Wikipedia abstract dataset using ColBERTv2 as a secondary, "best effort" source Khattab et al. (2023). This component is disabled by default to ensure the core application remains fully functional offline without need for large data storage requirements (1.4 GB English abstracts 2020). For a complete set of prompts for all agents, please refer to the Appendix. Relevant portions of the Appendix are called out in-line throughout.

2.3. Deployment and Reproducibility

A core objective of this work is to produce a fully *deployable and reproducible* tool. The Surgical Information Assistant is implemented in Python and is designed to run entirely offline on consumer-grade hardware. The key components include:

- **Language Models**: We use the Llama-3.2-3B model as the primary LLM for all language generation, selected for its balance of performance in medical QA datasets and computational efficiency Aaditya (2024); NVIDIA (2024). For comparison, we also evaluate against GPT-4o and Llama-3.1-8B via API calls where noted. All inference was performed with a temperature of 0.

- **API and User Interface**: We provide an API-level and user-friendly web interface built with Streamlit, allowing for easy interaction with the system.

- **Open Source**: The entire codebase, including the FAISS index, prompts, and evaluation scripts, will be released publicly to encourage further research and adaptation.

For large-scale testing on QA datasets, the OpenAI API was used to call GPT-4o (March-August 2025) and the TogetherAI API was used to evaluate all other models (March-August 2025). DeRetSyn with the Llama-3.2-3B model was also benchmarked for runtime performance on consumer hardware with 8 CPU cores (Intel i7 10th Gen) and 16GB of RAM (no GPUs) running both Ubuntu 24.04 and Windows 10, as well as, a M2 Macbook Pro with 32BM RAM and a M3 Macbook Air with 16GB RAM both running Sequoia 15.5. For all testing, we use the 8-bit quantized version for Llama-3.1-8B and Llama-3.2-3B. We also test the 4-bit quantized version of Llama-3.2-3B for runtime and memory footprint benchmarking. We use applications like LM Studio (2024) or Ollama (2023) for local testing (based on operating system support) which wrap llama.cpp and use memory mapping to load model shards into RAM Gerganov (2023). We compare these results with community-reported benchmarks of Llama-3.2-3B on mobile devices.

3. Evaluation Framework

Our evaluation was designed to assess the Surgical Information Assistant across three key dimensions: 1) the quality and reliability of the novel QA dataset; 2) the accuracy and robustness of the DeRetSyn workflow compared to established baselines; and 3) the system's practical performance on deployment-relevant metrics.

3.1. The OpenManualOfSurgeryQA Dataset

To facilitate reproducible research in low-resource clinical AI, we created and validated the *OpenManualOfSurgeryQA* dataset, a new dataset for surgical question-answering specific to resource-constrained settings.

Generation Process Starting with the OMSRS corpus, we divided the text into pseudo-randomly sized chunks (500 to 5000 characters) of varied context lengths and programmatically generated over 16,000 question-answer pairs using a Mistral Small 3.1 24B model to be answerable from these context chunks. The QA generation prompt consisted of several critical instructions. First, it mandates *groundedness*, requiring that answers be derived exclusively from the provided text. Second, it requires *self-containment*, ensuring questions are fully understandable without the source passage. Third, it directs the model to create questions that are *research-able* having enough context that relevant information and passages can be retrieved to answer the question. Finally, the prompt guides the model to generate a diversity of question types (e.g., definition, causality, comparison, yes/no) and to synthesize answers concisely rather than merely extracting sentences verbatim. Lastly, the prompt mandates a step-by-step reasoning process and a structured output format. We also include a 1-shot example to help encourage instruction-following and format consistency.

Quality Curation To ensure the dataset's quality and mitigate biases from automated generation, we implemented a multi-stage curation process. An initial manual review of 1,000 pairs by two individuals, including a practicing clinician, identified common failure modes (e.g., questions referencing figures or assuming external context). These patterns were used to create automated string-matching filters to remove low-quality pairs. Subsequently, each remaining QA pair was validated by an LLM against strict criteria for clarity, answerability, and faithfulness to the source text reducing the QA set further. Upon final manual review of another 1,000 pairs in the remaining set of 14,529, we did not discover any ambiguous or malformed QA pairs. This is the QA set we refer to as the *OpenManualOfSurgeryQA* dataset and use for further validation in this study.

3.2. Experimental Design

We conducted a series of experiments to answer three primary research questions:

1. **Efficiency vs. Scale**: Can a compact, offline model with an advanced retrieval workflow (DeRetSyn) outperform larger, state-of-the-art models that rely on parametric knowledge alone?
2. **Workflow Efficacy**: How does the multi-step DeRetSyn process compare against other RAG paradigms like vanilla retrieval and ReAct?
3. **Component Impact**: What are the specific contributions of the decomposition and fallback mechanisms to the system's overall performance?

Comparative Models and Ablations To answer these questions, we benchmarked several configurations. On our *OpenManualOfSurgeryQA* dataset, we compared DeRetSyn against baseline Llama-3.1-8B model, GPT-4o, vanilla RAG, and ReAct prompting Yao et al. (2023b) all with instructions for Chain-of-thought formatting. We also performed ablations by disabling the decomposition and fallback modules in DeRetSyn to isolate their impact.

To assess generalizability and mitigate the risk of evaluation circularity, we also evaluated DeRetSyn on the public *PubMedQA* dataset Jin et al. (2019) which contains "oracle" context for each question containing all relevant information. On this dataset, we compared our system against standard Chain-of-Thought (CoT) prompting Wei et al. (2022), both with and without provided context, to measure DeRetSyn's ability to utilize evidence and reason effectively.

3.3. Metrics and Validation

Automated Accuracy Assessment We measured accuracy using a *LLM-as-judge* (Mistral Small 3.1 24B). The judge was prompted to evaluate if a generated answer was factually consistent, complete, and relevant compared to the ground-truth answer. To ensure reliability, the prompt included three few-shot examples of correct and incorrect evaluations.

Human Validation of the LLM-as-judge The validity of using an LLM-as-judge is critical for the credibility of our results. We conducted a blinded human evaluation on 100 randomly sampled model outputs (50 judged correct, 50 incorrect). A human evaluator, using the same criteria, showed a high degree of agreement with the LLM-as-judge, achieving a Cohen's κ score of 0.90 (Table 1). This strong correlation provides confidence in our automated evaluation pipeline.

Table 1: Confusion matrix validating the LLM-as-judge against human evaluation for 100 QA pairs.

Human Label	LLM: Correct	LLM: Incorrect
Correct	48	3
Incorrect	2	47

We report 95% confidence intervals (CIs) for our accuracy metrics. Confidence intervals are computed either through the binomial approximation method for proportions of a population or through bootstrapping—we indicate which is used for all instances. To account for residual labeling uncertainty in the LLM-as-judge (Cohen's $\kappa = 0.90$ with human raters), we conservatively expand CIs by 10% to reflect the expected variance due to imperfect adjudication.

Deployment-Focused Benchmarking Beyond reporting accuracy on QA datasets, a primary goal of this work is practical utility in low-resource settings. Therefore, we also measured key deployment metrics for the DeRetSyn system. These metrics include:

- **End-to-End Latency**: The average time from query submission to final response.
- **Computational Cost**: The average number of tokens generated per query, which serves as a proxy for computational load.
- **Memory Footprint**: The peak RAM usage during inference to ensure the system can run on devices with limited memory.
- **Contextual Performance**: We report the performance of the system on those questions across all datasets that pertain to the most common operative procedures and pathologies encountered in resource-limited surgical settings, as described in Wong et al. (2014).

These practical benchmarks are essential for assessing the real-world viability of the Surgical Information Assistant.

4. Results

4.1. DeRetSyn Performance on Surgical QA

On the *OpenManualOfSurgeryQA* dataset, the DeRetSyn workflow with a 3B Llama model achieved a top-1 accuracy of 63.0%. As shown in Table 2, this result substantially outperforms both a baseline GPT-4o (42.5%) and a much larger 8B Llama model with vanilla RAG (52.5%). This ∼10 percentage point improvement over standard RAG demonstrates the impact of the multi-step decomposition and refinement process.

Impact of Workflow Components Removing the initial decomposition step reduced accuracy by nearly 4 points to 59.2%, highlighting its importance in structuring the retrieval process. Note that the CIs of the full system compared with the ablated initial decomposition are *not* overlapping (Table 2). Disabling the fallback mechanism had a smaller impact (62.1% accuracy with overlapping CIs), which was expected since the evaluation questions were sourced from the primary corpus. The fallback to Wikipedia

Table 2: Accuracy on the OpenManualOfSurgeryQA dataset. DeRetSyn with a 3B model surpasses larger models and alternative RAG strategies. CIs are computed from the binomial approximation.

Model Configuration	Accuracy [95% CI] (%)
Llama-3.1-8B (base)	28.7 [26.9, 30.4]
GPT-4o (base)	42.5 [40.7, 44.3]
Llama-3.1-8B + Vanilla RAG	52.5 [50.7, 54.3]
Llama-3.2-3B + Vanilla RAG	53.8 [52.0, 55.6]
Llama-3.2-3B + ReACT RAG	53.5 [51.7, 55.3]
Llama-3.2-3B + DeRetSyn (Ablated: Decomposition)	59.2 [57.4, 61.0]
Llama-3.2-3B + DeRetSyn (Ablated: Fallback)	62.1 [60.3, 63.9]
Llama-3.2-3B + DeRetSyn (Full System)	**63.0** [61.2, 64.8]

was triggered in only 5% of cases for the full system, confirming its role as a supplementary, rather than essential, component for in-domain queries.

Performance Stability A bootstrap analysis of accuracy from 200 subsets containing 30 samples each (Figure 2) shows that DeRetSyn has the highest mean accuracy but also exhibits stable performance with a standard deviation similar to other configurations. From bootstrapping, the performance of the full DeRetSyn system was significantly different from all configurations ($p < 0.001$, Unpaired t-tests with Bonferroni correction) except for DeRetSyn with ablated fallback.

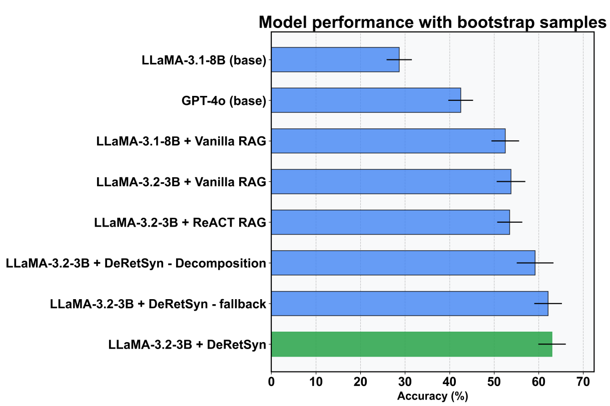

Figure 2: Bootstrapped accuracy comparison. DeRetSyn demonstrates higher mean accuracy and comparable variance to other configurations.

Computational cost We also computed the average number of tokens generated per response (Figure 3). Interestingly, Llama-3.2-3B generally used fewer tokens than the other LLMs tested in this study under similar configurations. However, the DeRetSyn system generates a larger number of total tokens than other RAG configurations due to the multiple generations required for each final response. Notably, the rate of increase in token generation is still less than expected when compared to the Llama-3.1-8B and GPT-4o base models under CoT prompting. While total generated tokens across all LLM calls for the full DeRetSyn system averaged ∼600, the final synthesized answers shown to users averaged ∼100 tokens.

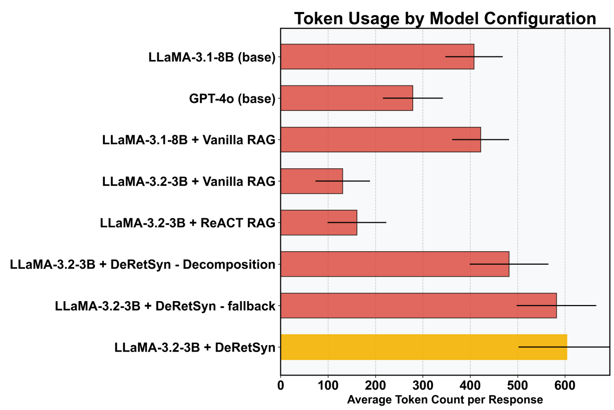

Figure 3: Total number of tokens generated from each configuration evaluated. Error bars represent the standard deviation in the dataset. Total tokens generated are not equal to the number of tokens in the final answer presented to the user due to CoT prompting and intermediate tool-usage steps.

4.2. Generalizability on the PubMedQA Dataset

To validate that DeRetSyn's effectiveness was not limited to our own dataset, we evaluated it on the public PubMedQA dataset. As shown in Table 3, DeRetSyn with a provided context achieves **70.1%** accuracy, outperforming both a similarly prompted GPT-4o (64.1%) and standard CoT prompting (63.5%). This demonstrates DeRetSyn's superior ability to utilize provided evidence for reasoning.

Notably, even when forced to rely only on its internal reasoning structure (*no context, no retrieval*), DeRetSyn (52.1%) still outperforms a baseline CoT model (50.0%), suggesting the decomposition process itself aids reasoning.

4.3. Practical Performance for Real-World Deployment

Beyond theoretical accuracy, the practical viability of the Surgical Information Assistant depends on its performance on consumer-grade hardware and its ability to support the most common operative procedures and pathologies encountered in resource-limited settings.

Memory Footprint and Feasibility on Consumer Hardware Memory consumption is an important factor for hardware deployment in low-resource settings. We benchmarked the Surgical Information Assistant on a range of consumer-grade hardware to measure its peak RAM usage during inference. As shown in Table 4, the peak memory usage for the Llama-3.2-3B model was consistently under *3.5 GB* across all platforms. This is lower than expected due to memory mapping that is handled by llama cpp by only loading necessary parts of the model into memory. This low memory footprint confirms that the entire system can run on standard laptops and devices with as little as 8 GB of total RAM. The results also show a modest latency improvement and lower memory consumption with 4-bit quantization.

Performance on Relevant Procedures and Pathologies We automatically classified all questions in the *PubMedQA* and *OpenManualOfSurgeryQA* datasets with zero-shot classification using the `bart-large-mnli` model to determine which ones were relevant to the most common operative procedures and pathologies in resource-limited settings as reported in Wong et al. (2014). These included 'Cesarean delivery', 'herniorrhaphy', 'wound debridement and care', 'percutaneous infections and abscesses', and 'circumcision'. We also included an 'other' category for those questions that did not pertain to any of the above. We found a total of 1372 questions across all relevant categories. The DeRetSyn system had a higher accuracy for this subset of questions (75.1%, 96% CI: [72.7, 77.3]) than for the entire *OpenManualOfSurgeryQA* data.

4.4. Analysis of Failure Modes

A qualitative review of incorrect answers revealed three primary failure modes:

- **Sub-question Misalignment**: Occurs when the decomposition agent misinterprets the intention of a query, leading to sub-questions that are too broad or tangential.
- **Synthesis Drift**: In cases of sparse or conflicting retrieved evidence from nuanced questions (e.g., PubMedQA 'Maybe' items), the synthesis agent occasionally blended factual information with plausible-sounding fabrications. These errors are usually visible in intermediate steps, supporting safe human oversight.
- **Generic Fallback**: When the fallback to Wikipedia was triggered, the retrieved information sometimes lacked the necessary clinical specificity.

These findings underscore the need for ongoing work in prompt engineering for clinical nuance, improving synthesis grounding, and curating more domain-specific fallback resources.

5. Discussion

Our findings demonstrate that with the an appropriate architecture, compact, locally-runnable models can outperform cloud-based systems in specialized domains. This work provides a practical blueprint for developing LLM-based tools that are transparent and deployable. While the system described clearly has limitations, we believe it presents an architecture that aligns with the technical infrastructure requirements of a resource-constrained clinical setting and improves over baselines and much larger models without extensive fine-tuning. Each component of this architecture can be further improved while staying aligned with the core requirements for a resource-limited setting.

5.1. Bridging the Performance Gap with Efficient Orchestration

A central result of our study is that knowledge access and intelligent workflow orchestration are more critical than raw model scale for specialized tasks. The Llama-3.2-3B model, when augmented with the DeRetSyn workflow, surpassed the much larger GPT-4o and a vanilla RAG implementation with an 8B model for the datasets we tested. This confirms that the performance gap between large and small models can be effectively closed by shifting focus from scaling parametric knowledge to optimizing the retrieval and synthesis process Singhal and et al. (2023); Wang and et al. (2024).

The DeRetSyn method's 10.5 percentage point accuracy gain over standard RAG highlights the value of

Table 3: Performance on the PubMedQA dataset. DeRetSyn demonstrates superior reasoning and evidence utilization compared to baselines. CIs are computed from the binomial approximation.

Method	Overall [95% CI] (%)	'Yes' (%)	'No' (%)	'Maybe'(%)
1. LLM-only w/o CoT and w/o context	48.50 [45.4, 51.7]	46.92	63.31	10.91
2. LLM-only w/ CoT and w/o context	50.00 [46.9, 53.2]	66.12	35.80	12.73
3. LLM-only w/ CoT and fixed PubMedQA context	63.50 [60.4, 66.4]	71.56	66.57	13.64
4. DeRetSyn (no context, no fallback, no retrieval)	52.10 [49.0, 55.2]	67.39	39.94	12.73
5. DeRetSyn w/ fixed PubMedQA context (no fallback, no retrieval)	**70.10** [67.2, 72.9]	**79.89**	**71.89**	15.45
6. DeRetSyn w/ OMSRS retrieval and fallback (no PubMedQA context)	61.50 [58.4, 64.5]	68.84	65.09	13.63
7. GPT-4o CoT w/ fixed PubMedQA context	64.10 [61.0, 67.1]	60.87	68.05	**68.18**

Table 4: Performance and memory benchmarks on consumer hardware of the DeRetSyn-powered Llama-3.2-3B model.

Hardware Platform	Quantization	Avg. Latency (s)	Peak RAM (GB)
Windows/Ubuntu (8 CPU, 16GB RAM)	8-bit	14.2 - 15.6	3.1 - 3.2
	4-bit	13.5 - 13.8	2.5
M2 Macbook Pro (32GB RAM)	8-bit	8.2	2.6
	4-bit	7.7	2.2
M3 Macbook Pro (16GB RAM)	8-bit	8.1	2.5
	4-bit	7.7	2.1

its structured, multi-step approach. By decomposing complex queries, the system emulates a more deliberate reasoning process, which, as our PubMedQA results show, improves both the utilization of provided context and the model's intrinsic reasoning capabilities.

5.2. Implications for Health Equity and Responsible AI

The practical impact of this work is rooted in its potential to advance health equity. The Surgical Information Assistant is designed to be fully operational on consumer-grade hardware without an internet connection, directly addressing the digital divide that excludes many low-resource clinical settings from the benefits of modern AI. A preliminary analysis shows that the system handles queries about common procedures and pathologies for resource-limited settings with a higher accuracy than most queries overall. However, this requires further investigation within more clinically relevant settings. Critically, the current performance of the system (∼60%–70% accuracy) is likely not adequate for immediate use by clinicians in the field. We discuss how this can be improved further.

Trust and Interpretability in a Clinical Workflow Beyond accessibility, the DeRetSyn architecture promotes trust and transparency. DeRetSyn's modularity makes the reasoning process explicit; by surfacing the intermediate sub-questions, retrieved evidence, and synthesized sub-answers, it allows a clinician to quickly verify the basis of a recommendation. Moreover, our system can localize patient data since it can run on-device enabling data security. This aligns with principles of interpretable AI Wang et al. (2022); Yao et al. (2023a) and is essential for safe and responsible deployment in medicine Zhao and et al. (2024). In future studies, we plan to leverage this transparent reasoning to study how this tool can improve clinical judgment.

The Accuracy vs. Efficiency Tradeoff DeRetSyn's accuracy comes at the cost of increased token generation (Figure 3). However, our benchmarks on laptops confirm that this latency remains well within practical limits. On mobile devices, the Llama-3.2-3B model with a llama cpp deployment generates 17–23 tokens/second Arm Newsroom (2024); Qualcomm (2024); Hotellnx (2024); ctrl-brk (2024). Assuming an average of 600 tokens for the final response (Figure 3), latency per query is estimated at ∼30 seconds. Devices capable of running multiple model instances in parallel may reduce this time through concurrent execution. Moreover, user interactions with the intermediate steps of the system may be important for transparency and auditability, making the total latency not the sole focus of usability.

5.3. Limitations and Future Directions

While our results are promising, we identify many areas for future work:

Dataset and Evaluation: The primary evaluation relied on a dataset generated from the source corpus. Although a strong performance on PubMedQA mitigates concerns of circularity, future work should validate the system against a dataset created from real-world clinical queries. Furthermore, future QA evaluations should incorporate more complex, multi-hop reasoning scenarios.

Prospective Clinical Evaluation: The ultimate test of this system is its performance in a real clinical setting. We are actively planning a prospective study to evaluate the Surgical Information Assistant with clinicians in a simulated low-resource environment. This will be crucial for assessing usability, measuring its impact on decision-making, and identifying unforeseen failure modes or risks.

Human-in-the-Loop Refinement: The current DeRetSyn workflow is fully automated. A promising future direction is to incorporate human-in-the-loop capabilities, allowing a clinician to correct a flawed sub-question or provide additional context during the refinement process. This could further enhance accuracy and usability.

Multi-language support: While DeRetSyn is language-agnostic, the OMSRS corpus is in English which constrains the current language of the reference documentation. We tested a 100-question subset of the *OpenManualOfSurgeryQA* translated to Spanish, showing comparable accuracy after back-translation of the final answer with confidence intervals overlapping with of the full system (-1.5% drift of mean accuracy). Future work will investigate this further and integrate multilingual front-ends and bilingual retrieval indexes for broader accessibility.

Knowledge and confidence: While DeRetSyn achieves strong results relative to the size of the LLM, its reasoning and retrieval accuracy are ultimately bounded by the pretrained model's latent knowledge and its ability to assess uncertainty. Hypothetically, both could be addressed through fine-tuning and confidence modeling efforts. Specifically, curating "good" DeRetSyn traces from multiple medical QA datasets could be used to finetune smaller LLMs for improved domain adaptation and reasoning stability. In parallel, developing internal confidence estimators at each sub-question and synthesis stage to quantify uncertainty and surface low-confidence outputs to users could improve transparency and performance. Together, these efforts could strengthen both the factual grounding and epistemic transparency of the system.

Potential harms: Potential harms include over-confident or outdated recommendations, especially under sparse evidence. The system can partially mitigate this through explicit deferrals ('insufficient evidence') and the visible citations per sub-question. Future work includes uncertainty-aware confidence scores and clinician-in-the-loop verification during the initial decomposition.

6. Conclusion

We show how the Surgical Information Assistant addresses resource-constraints and transparency by ensuring offline access to relevant medical information through explicit iterative reasoning and synthesis with reasonable latency.

Acknowledgments

We acknowledge Dr. Sankha Banerjee for their support in helping to review question-answer pairs and system responses.

References

K Aaditya. Performance Comparison: Llama-3.2 vs. Llama-3.1 LLMs and Smaller Models (3B, 1B) in Medical and Healthcare AI Domains. Hugging Face Blog, Sep 2024. URL https://huggingface.co/blog/aaditya/llama3-in-medical-domain.

BillSmith Anyinkeng Achanga, Christian Wabene Bisimwa, Victor Oluwafemi Femi-Lawal, Nnoko Sona Akwo, and Tohson Falake Toh. Surgical practice in resource-limited settings: Perspectives of medical students and early career doctors: A narrative review. *Health Science Reports*, 8(1):e70352, 2025.

Arm Newsroom. Ai inference everywhere with new llama llms on arm. https://newsroom.arm.com/news/ai-inference-everywhere-with-new-llama-llms, 2024. URL https://newsroom.arm.com/news/

ai-inference-everywhere-with-new-llama-llms. Accessed: 2025-06-06.

David J Bunnell, Mary J Bondy, Lucy M Fromtling, Emilie Ludeman, and Krishnaj Gourab. Bridging ai and healthcare: A scoping review of retrieval-augmented generation—ethics, bias, transparency, improvements, and applications. *medRxiv*, pages 2025–04, 2025.

ctrl-brk. Phone LLM's benchmarks? Reddit, Nov 2024. URL https://www.reddit.com/r/LocalLLaMA/comments/1glx6a5/phone_llms_benchmarks/.

Georgi Gerganov. llama.cpp. https://github.com/ggml-org/llama.cpp, 2023.

Hotellnx. Run Llama 3.2 3B on Phone - on iOS & Android. Reddit, Oct 2024. URL https://www.reddit.com/r/LocalLLaMA/comments/1fppt99/run_llama_32_3b_on_phone_on_ios_android/.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2201–2211, 2019.

Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187, 2025.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Siru Liu, Allison B McCoy, and Adam Wright. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, page ocaf008, 2025.

LM Studio. Lm studio documentation, 2024. URL https://lmstudio.ai/docs/app. Accessed: 2024-08-22.

Karen Ka Yan Ng, Izuki Matsuba, and Peter Cheng-ming Zhang. Rag in health care: a novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*, 2 (1):AIra2400380, 2025.

NVIDIA. Llama-3.2-3b-instruct Model by Meta - NVIDIA NIM APIs. https://build.nvidia.com/meta/llama-3.2-3b-instruct/modelcard, 2024. Accessed: August 4, 2025.

Ollama. Ollama. https://github.com/ollama/ollama, 2023.

Qualcomm. qualcomm/Llama-v3.2-3B-Instruct. Hugging Face, Sep 2024. URL https://huggingface.co/qualcomm/Llama-v3.2-3B-Instruct.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Xiang Li, and Ninghao Liu. Mkrag: Medical knowledge retrieval augmented generation for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2024, page 1011, 2025.

K Singhal and et al. Retrieval-augmented language models for clinical medicine. *Nature*, 620:282–289, 2023.

Vanderbilt University Medical Center. About the global surgical atlas. https://www.vumc.org/global-surgical-atlas/about, 2025. Accessed: 2025-04-02.

Josip Vrdoljak, Zvonimir Boban, Marino Vilović, Marko Kumrić, and Joško Božić. A review of large language models in medical education, clinical decision support, and healthcare administration. In *Healthcare*, volume 13, page 603. MDPI, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Z Wang and et al. Medgraphrag: Safe medical llm via graph-based retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Evan G Wong, Miguel Trelles, Lynette Dominguez, Shailvi Gupta, Gilbert Burnham, and Adam L Kushner. Surgical skills needed for humanitarian missions in resource-limited settings: common operative procedures performed at medecins sans frontieres facilities. *Surgery*, 156(3):642–649, 2014.

Kehan Xu, Kun Zhang, Jingyuan Li, Wei Huang, and Yuanzhuo Wang. Crp-rag: A retrieval-augmented generation framework for supporting complex logical reasoning and knowledge planning. *Electronics*, 14(1):47, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. *URL https://arxiv. org/abs/2305.10601*, 3:1, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Hao Zhang and et al. Discuss-rag: Enhancing retrieval-augmented generation via agent-led multi-turn reasoning for medical qa. *arXiv preprint arXiv:2504.21252*, 2024.

X Zhao and et al. Aipatient: Simulating patients with ehrs and agentic workflow. *arXiv preprint arXiv:2409.18924*, 2024.

Appendix A. Instruction prompt for QA generation from text fragments

```
You are an expert medical curriculum developer. Given a medical or medically-related text passage, you must generate question-answer pairs. The questions MUST be
- self-contained: answerable by the information in the passge but understandable on their own without needing direct reference to the passage (ie. "what is this passage about?" is NOT a valid question).
- grounded: directly related to the information in the passage
- research-able: have enough information to be able to retrieve the relevant passages or information in the future to help answer the question.
Generate a set of question-ansewr types including yes/no answer, short phrases, and longer sentences.

Think step-by-step and reason through the content of the passage to hypothesize potential question-answer pairs and then respond. Think of at least 8 question-answer pairs but more, if possible.

Respond in this format:
<think> You reasoning here... </think>
<qa-pair> [question here...] | [answer here...] </qa-pair>
<qa-pair> [question here...] | [answer here...] </qa-pair>

Here is an example:

PASSAGE:
The goal of transvesical prostatectomy is to
remove the hypertrophic transitional zone of the
prostate to reduce urinary obstruction. It is not a
surgical procedure for malignant lesions.
Appropriate history (including functional
```

status inquiry) and physical examination are done
prior to any investigation. Physical exam always
includes digital rectal exam. Enlarged lateral lobes
can be easily felt and the rectal mucosa should move
freely over the prostate. Any nodularity or fixation of
the rectal mucosa should alert you to the possibility
of prostatic cancer and core needle biopsy should be
done before proceeding with the admission.

RESPONSE:
<think>
Okay, so I need to generate question-answer pairs based on the given medical passage. The questions should be answerable from the passage and understandable on their own. Let me read through the passage carefully to understand the key points.

The passage is about transvesical prostatectomy. It mentions that the goal is to remove the hypertrophic transitional zone of the prostate to reduce urinary obstruction. It's not for malignant lesions. Before any investigation, they do a history and physical exam, including a digital rectal exam. Enlarged lateral lobes can be felt, and the rectal mucosa should move freely. If there's nodularity or fixation, it might indicate cancer, so a biopsy is needed before admission.

Alright, so the main topics here are the purpose of the procedure, what it's not used for, the pre-procedure assessments, the physical exam findings, and when to suspect cancer.

I should think of questions that cover these areas. Let's start with the purpose. A question like, "What is the primary goal of transvesical prostatectomy?" The answer would be about removing the hypertrophic zone to relieve obstruction.

Next, since it's not for malignant lesions, a question could be, "Is transvesical prostatectomy used to treat malignant lesions of the prostate?" The answer would be no.

Then, the pre-procedure steps. Maybe, "What assessments are typically performed before transvesical prostatectomy?" The answer would include history, physical exam, and digital rectal exam.

For the physical exam findings, a question could be, "What is a normal finding during the digital rectal exam in a patient undergoing transvesical prostatectomy?" The answer would mention enlarged lateral lobes and freely moving rectal mucosa.

When to suspect cancer: "What signs during a digital rectal exam may indicate the need for a core needle biopsy?" The answer would be nodularity or fixation of the rectal mucosa.

I can also think about the implications of the findings. For example, "What should be done if nodularity or fixation is detected during the digital rectal exam?" The answer would be to perform a biopsy before proceeding.

Another question could be about the importance of the digital rectal exam in this context. "Why is a digital rectal exam important before transvesical prostatectomy?" The answer would be to assess for signs of cancer.

I should make sure each question is clear and can be answered without directly referencing the passage, just the information in it. Let me see if I can come up with more. Maybe about the specific zones of the prostate involved. "Which part of the prostate is targeted in transvesical prostatectomy?" The answer is the transitional zone.

Also, the procedure's focus: "What is the focus of transvesical prostatectomy in treating urinary obstruction?" The answer is removing the hypertrophic transitional zone.

I think that's a good set. I have eight pairs, but maybe I can find a couple more. Perhaps about the consequences of not doing the biopsy. "What is the consequence of not performing a core needle biopsy when suspicious findings are present?" The answer would be proceeding without knowing if cancer is present, which could be risky.

Wait, the passage doesn't explicitly state the consequences, so maybe that's beyond the given info. I should stick to what's in the passage.

Another angle: the patient's history. "What is included in the patient's history before transvesical prostatectomy?" The answer would be functional status inquiry.

I think that's a solid list. I'll make sure each question is clear and each answer is accurate based on the passage.

</think>

<qa-pair> What is the primary goal of transvesical prostatectomy? | The primary goal is to remove the hypertrophic transitional zone of the prostate to reduce urinary obstruction. </qa-pair>

<qa-pair> Is transvesical prostatectomy used to treat malignant lesions of the prostate? | No, it is not a surgical procedure for malignant lesions. </qa-pair>

<qa-pair> What assessments are typically performed before transvesical prostatectomy? | Appropriate history, including functional status inquiry, and physical examination are done prior to any investigation. </qa-pair>

<qa-pair> What is a normal finding during the digital rectal exam in a patient undergoing transvesical prostatectomy? | Enlarged lateral lobes can be easily felt, and the rectal mucosa should move freely over the prostate. </qa-pair>

<qa-pair> What signs during a digital rectal exam may indicate the need for a core needle biopsy? | Nodularity or fixation of the rectal mucosa may indicate the possibility of prostatic cancer. </qa-pair>

<qa-pair> What should be done if nodularity or fixation is detected during the digital rectal exam? | A core needle biopsy should be performed before proceeding with the admission. </qa-pair>

<qa-pair> Why is a digital rectal exam important before transvesical prostatectomy? | It is important to assess for signs of prostatic cancer, such as nodularity or fixation of the rectal mucosa. </qa-pair>

<qa-pair> Which part of the prostate is targeted in transvesical prostatectomy? | The transitional zone of the prostate is targeted. </qa-pair>

<qa-pair> What is the focus of transvesical prostatectomy in treating urinary obstruction? | The focus is on removing the hypertrophic transitional zone of the prostate. </qa-pair>

<qa-pair> What is included in the patient's history before transvesical prostatectomy? | The patient's history includes a functional status inquiry. </qa-pair>

Here is the passage for which you need to generate question-answer pairs

PASSAGE:
{text_chunk}

RESPONSE:

Appendix B. System prompt for LLM-as-judge

ROLE AND OBJECTIVE

You are an impartial and meticulous AI evaluator. Your objective is to determine if a "Generated Answer" correctly and completely answers a given "Query", using the "Ground Truth Answer" as the definitive source of

```
correctness. Your evaluation must result in a binary decision: "Correct" or "Incorrect".

### EVALUATION CRITERIA ###
You must adhere to the following strict criteria:

1. **Factual Consistency:** The "Generated Answer" must be factually consistent with the "Ground Truth Answer". It must not contain any information that contradicts the ground truth.
2. **Completeness:** The "Generated Answer" must address all parts of the "Query". It is considered "Incorrect" if it omits critical information that is present in the "Ground Truth Answer" and is necessary for a full response.
3. **Relevance:** The "Generated Answer" must directly answer the user's "Query". An answer that is factually correct but irrelevant to the question is "Incorrect".

**Important Note:** Differences in phrasing, verbosity, or style between the "Generated Answer" and the "Ground Truth Answer" are acceptable as long as the core semantic meaning is the same and the criteria above are met.
```

B.1. Instruction prompt prefix for LLM-as-judge

```
### INSTRUCTIONS ###
Follow these steps to generate your evaluation:

1. **Analyze the Query:** Carefully read the user's "Query" to understand the user's full intent and all a-priori information.
2. **Study the Ground Truth:** Read the "Ground Truth Answer" to understand the complete and factually correct response. This is your gold standard.
3. **Compare:** Compare the "Generated Answer" against the "Ground Truth Answer" based on the **Factual Consistency** and **Completeness** criteria.
4. **Final Judgment:** Based on your comparison, make a final binary judgment.
5. **Output Format:** Provide your response in the following format
<think> Your reasoning here... </think>
<answer> correct OR incorrect </answer>

Here are some examples

**Example 1**

QUERY
What is the escape velocity from the surface of Earth?

GROUND TRUTH ANSWER
The escape velocity from Earth's surface is approximately 11.2 kilometers per second (km/s), which is about 40,320 kilometers per hour (km/h) or 25,000 miles per hour (mph).

GENERATED ANSWER
To escape Earth's gravity from the surface, an object needs to reach a speed of about 11.2 km/s.

YOUR RESPONSE
<think> Step 1: The query asks for Earth's escape velocity. Step 2: The ground truth states it is 11.2 km/s. Step 3: The generated answer also states it is 11.2 km/s. Step 4: The generated answer is factually consistent and directly answers the query. While less detailed than the ground truth, it does not omit the critical piece of information and contains no contradictions. Therefore, it is correct. </think>
<answer> correct </answer>

**Example 2**

QUERY
Who was the second person to walk on the moon?
```

```
GROUND TRUTH ANSWER
Neil Armstrong was the first; Buzz Aldrin
    was second.

GENERATED ANSWER
The second man on the moon was Michael
    Collins.

YOUR RESPONSE
<think> Step 1: The query asks for the
    second person on the Moon. Step 2: The
    ground truth identifies this person as
    Buzz Aldrin. Step 3: The generated
    answer incorrectly identifies the
    person as Michael Collins. Step 4:
    This is a direct factual contradiction
    with the ground truth. Therefore, the
    answer is incorrect. </think>
<answer> incorrect </answer>

**Example 3**

QUERY
What are Newton’s first two laws of motion?

GROUND TRUTH ANSWER
Newton’s first law states that an object
    will not change its motion unless a
    force acts on it. The second law
    states that the force on an object is
    equal to its mass times its
    acceleration.

GENERATED ANSWER
Newton’s first law is the law of inertia,
    stating an object in motion stays in
    motion.

YOUR RESPONSE
<think> Step 1: The query asks for
    Newton’s first AND second laws. Step
    2: The ground truth provides both
    laws. Step 3: The generated answer
    only provides the first law. Step 4:
    The generated answer is incomplete as
    it omits a critical part of the
    information required by the query and
    present in the ground truth.
    Therefore, it is incorrect. </think>
<answer> incorrect </answer>

Here are the "Query", the "Ground Truth
    Answer", and the "Generated Answer".

QUERY
{query}

GROUND TRUTH ANSWER
{ground_truth_answer}

GENERATED ANSWER
{generated_answer}

YOUR RESPONSE
```

Appendix C. Prompts Used in the Surgical Information Assistant

This appendix contains all the prompts used in the Surgical Information Assistant codebase. These prompts are used to guide the language models in performing various tasks such as question decomposition, information retrieval, answer synthesis, searching a fallback data-store, best-effort answer generation, and follow-up question generation.

C.1. Question Decomposition Prompt

This prompt is used to break down complex surgical questions into simpler sub-questions:

```
You are an expert at breaking complex
    surgical questions into simpler ones.
    Break the following question into
    smaller sub-questions:

Question: {question}

Each sub-question should be independent
    and answerable on it’s own without
    needing reference to other
    sub-questions. Think of at least 3
    sub-questions but no more than 7.

Think step-by-step and make sure to reason
    through how break the question into
    sub-questions.
```

```
Create new sub-questions in the following format but do NOT answer the question. Respond in the following format:

<think> Your reasoning here... </think>
<sub-question> The first sub-question... </sub-question>
<sub-question> The second sub-question... </sub-question>
...
<sub-question> The last sub-question... </sub-question>
```

C.2. Answer Generation from Context Prompt

This prompt is used to generate answers from retrieved context for each sub-question. This results in multiple LLM calls, but can be run asynchronously:

```
Based on the given question and context, generate an answer.
Question: {question}
Context: {context}

Think step-by-step and make sure to reason through how to generate an answer. ONLY rely on the given context to generate the answer.

Include snippets of the context that support your answer. Do NOT use any information outside of the given context to generate the answer.

Respond in the following format:

<think> Your reasoning here... </think>
<answer> The generated answer... </answer>
<snippet> First relevant snippet from the context... </snippet>
<snippet> Second relevant snippet from the context... </snippet>
...
<snippet> The last relevant snippet from the context </snippet>
```

C.3. Prompt for answer generation without context

This prompt is used to generate answers for sub-questions when no context is available (for ablation experiments).

```
You are a medical expert specializing in surgery. Answer the following question using your knowledge of surgical procedures, anatomy, and medical practices.

Question: {question}

Think step-by-step and provide a comprehensive answer based on your medical knowledge. If you're uncertain about any aspect, please indicate that in your response.

Respond in the following format:

<think> Your reasoning here... </think>
<answer> The generated answer based on your medical knowledge... </answer>
<confidence> High/Medium/Low - your confidence level in this answer </confidence>
```

C.4. Answer Synthesis Prompt

This prompt is used to synthesize answers from multiple sub-questions or route for further iteration in the DeRetSyn system.

```
You are a reasoning engine. Given the following sub-question answers, determine whether they are enough to fully answer the original question. ONLY rely on the knowledge to determine whether the question can be answered.

If yes, then provide the answer. Make your answer detailed and structured with sections, as appropriate. Include as much relevant information as possible from the knowledge provided.

If you determine that you cannot answer the original question, then suggest
```

```
    what additional questions should be
    asked.

Original Question:
{original_question}

Knowledge:
{answers}

Think step-by-step to reason through you
    answer and consider the relevant
    information. Respond in the following
    format:
<think> Your reasoning here... </think>
<can_answer> yes OR no </can_answer>
<answer> The answer to the original
    question... </answer>
<new_questions> The first new
    sub-question... </new_questions>
<new_questions> The second new
    sub-question... </new_questions>
...
<new questions> The last new sub-question
    </new_questions>
```

C.5. Wikipedia Search Prompt

This prompt is used to generate search queries for Wikipedia:

```
Given the following question, generate 3
    search queries that would help find
    relevant information on Wikipedia. The
    queries should be specific and focused
    on the key concepts in the question.

Question: {question}

Respond in the following format:
<query>first search query</query>
<query>second search query</query>
<query>third search query</query>
```

C.6. Wikipedia Context Synthesis Prompt

This prompt is used to synthesize information from Wikipedia contexts:

```
You are a reasoning engine. Given the
    following question and Wikipedia
    contexts, synthesize the information
    to provide a comprehensive answer.

Question: {question}

Wikipedia Contexts:
{contexts}

Think step-by-step to reason through your
    answer and consider the relevant
    information from the contexts. Respond
    in the following format:
<think> Your reasoning here... </think>
<answer> The synthesized answer...
    </answer>
```

C.7. Best Effort Answer Generation Prompt

This prompt is used when the system needs to generate a best-effort answer using Wikipedia:

```
You are a reasoning engine. Given the
    following original question and
    sub-question answers, formulate an
    answer to the best of your ability.

Original Question:
{original_question}

Knowledge:
{state["answers"]}
{state["wikipedia_results"]}

Think step-by-step to reason through you
    answer and consider the relevant
    information. Respond in the following
    format:
<think> Your reasoning here... </think>
<answer> The answer to the original
    question... </answer>
```

C.8. Follow-up Question Generation Prompt

This prompt is used to generate follow-up questions:

```
You are a reasoning engine. Given the
    following original question and final
    answer, generate 3 follow-up questions
    that help expand on the original
    question and the answer in a step-wise
    manner.

Original Question:
{original_question}
```

```
Final Answer:
{final_answer}

Think step-by-step to reason through your
    answer and consider the relevant
    information. Respond in the following
    format:
<think> Your reasoning here... </think>
<follow_up_questions> follow-up question
    here... </follow_up_questions>
<follow_up_questions> follow-up question
    here... </follow_up_questions>
<follow_up_questions> follow-up question
    here... </follow_up_questions>
```

C.9. Chain-of-Thought Generation Prompt

This prompt is used to generate detailed reasoning for answers:

```
You are a reasoning engine. Based on the
    following question and knowledge,
    provide a detailed, step-by-step
    reasoning to arrive at an answer.
    Include at least 3 steps, but more as
    needed.

Question:
{state["original_question"]}

Knowledge:
{state["answers"]}
{state["wikipedia_results"] if
    "wikipedia_results" in state else ""}

Provide your response in this format:

<think> Your reasoning here... </think>
<answer> The final answer here... </answer>
```

Appendix D. Chat interaction management prompts

The following prompts are used to manage the chat-interface with the Surgical Information Assistant.

D.1. Surgery Topic Classification Prompt

This prompt is used to determine if a question is about surgery. If so, the DeRetSyn system is triggered. If not, the LLM answers with general knowledge acknowledging it may hallucinate and directs the user to ask a more relevant question:

```
Determine if the following question
    requires access to specific medical
    documents related to detailed surgical
    information to be answered accurately.
Think step-by-step and reason through your
    answer. Respond in the following
    format:

<thinking> Your reasoning here...
    </thinking>
<answer> yes OR no </answer>

Here are some examples:

Question:
What is the primary purpose of the
    coronary artery bypass graft?
Response:
<thinking> The question is asking about a
    coronary bypass graft which is related
    to surgery. So yes, this question is
    about surgery. </thinking>
<answer> yes </answer>

Question:
Is machine learning useful for solving
    complex medical problems?
Response:
<thinking> The question is asking about
    machine learning as a tool for solving
    complex medical problems. While the
    question does specify medical
    problems, it does not refer to surgery
    or topics related to surgery.
    </thinking>
<answer> no </answer>

Question:
Can you elaborate on your previous
    response about suturing or rephrase it?
Response:
<thinking> The question is asking for an
    explanation of suturing that was given
    earlier in the conversation. While the
    question does mention suturing which
    is related to surgery, it is not
    asking about suturing specifically but
```

```
    rather requesting to explain a
    previous response. </thinking>
<answer> no </answer>

Question:
What are some core differences between
    robotic and laparasopic inguinal
    hernia?
Response:
<thinking> The question is asking about
    the differences between robotic and
    laparosopic inguinal hernia. Both
    robotic and laparosopic surgery
    involve surgery on the inguinal
    hernia. This is clearly a question
    about surgery and requires
    investigation into documentation to
    answer. </thinking>
<answer> yes </answer>

Question:
{question}
Response:
```

D.2. Context-Aware Input Update Prompt

This prompt is used to update user input with context from chat history:

```
Given the following chat history and a new
    user input, update the user input to
    include any relevant context from the
    chat history. Only add necessary
    context, and keep the updated input
    concise.
Think step-by-step and reason through your
    answer. Respond in the following
    format:

<thinking> Your reasoning here...
    </thinking>
<updated_input> Updated user input here...
    </updated_input>

Here are some examples:

CHAT HISTORY:
User: What are the main types of sutures
    used in surgery?
Assistant: The main types of sutures used
    in surgery are absorbable and
    non-absorbable sutures. Absorbable
    sutures dissolve over time, while
    non-absorbable sutures need to be
    removed. Common absorbable sutures
    include catgut and synthetic materials
    like polyglycolic acid. Non-absorbable
    sutures are often made of materials
    like nylon or silk.
User: How long do absorbable sutures
    typically last?
Assistant: The duration of absorbable
    sutures varies depending on the
    material used. Generally, they can
    last anywhere from a few days to
    several months. For example,
    fast-absorbing sutures may dissolve
    within 7-10 days, while
    slower-absorbing sutures can take
    60-90 days or even longer.
NEW USER INPUT:
What about non-absorbable ones?
RESPONSE:
<thinking> Considering the previous
    discussion the user is referring to
    non-absorbable sutures. Specifically,
    they are asking about the lifespan or
    removal time-frame for non-absorbable
    sutures. I'll restate the question to
    include this context from the chat
    history. </thinking>
<updated_input> What is the typical
    lifespan or removal timeframe for
    non-absorbable sutures used in
    surgery? </updated_input>



CHAT HISTORY:
User: What is a laparotomy?
Assistant: A laparotomy is a surgical
    procedure that involves making a large
    incision in the abdomen to gain access
    to the abdominal cavity. It allows
    surgeons to examine the abdominal
    organs, diagnose problems, and perform
    necessary treatments. Laparotomies are
    used for various purposes, including
    exploratory surgery, treatment of
    abdominal trauma, and removal of
    tumors.
```

```
User: Are there any alternatives to this procedure?
Assistant: Yes, there are alternatives to laparotomy, particularly minimally invasive techniques. The main alternative is laparoscopy, also known as keyhole surgery. In laparoscopy, several small incisions are made instead of one large incision. A camera and specialized instruments are inserted through these small incisions to perform the surgery. This technique often results in less pain, faster recovery, and smaller scars compared to traditional laparotomy.

NEW USER INPUT:
What are the risks?

RESPONSE:
<thinking> The user is asking about risks, but it's not clear whether they're referring to laparotomy or laparoscopy, both of which were discussed in the previous messages. Since laparotomy was the original topic and laparoscopy was introduced as an alternative, it would be helpful to ask about the risks of both procedures for a comprehensive answer.
</thinking>
<updated_input> What are the risks associated with both laparotomy and laparoscopy procedures?
</updated_input>


CHAT HISTORY:
User: What are the potential complications of laparoscopic surgery?
Assistant: Laparoscopic surgery can have several potential complications, including:
1. Infection: Laparoscopic surgery can lead to infections, especially in patients with infections that can be spread through the abdominal cavity.
2. Stool bleeding: Laparoscopic surgery can lead to stool bleeding, especially if the surgical instrument is inserted directly into the abdominal cavity.
3. Pain: Laparoscopic surgery can be painful, especially for patients with pre-existing conditions or those who have had previous laparoscopic surgeries.

NEW USER INPUT:
Can you think of any more?

RESPONSE:
<thinking> The user is asking if there are any more complications, but it's important to note that the original question was about laparoscopic surgery, not about potential complications. I'll restate the question to clarify that the original topic was laparoscopic surgery.
</thinking>
<updated_input> Are there any more potential complications associated with laparoscopic surgery besides infection, stool bleeding, and pain?
</updated_input>


CHAT HISTORY:
{formatted_history}

New User Input:
{user_input}

RESPONSE:
```

Appendix E. Prompts used to test alternative RAG paradigms

E.1. Chain-of-thought prompting without context

```
You are a medical expert. Please answer the following question:
Think step-by-step and provide a detailed reasoning process to arrive at your answer. Include at least 3 steps in your reasoning, but more as needed.
```

```
Respond in the following format:

<think> Your reasoning here... </think>
<answer> Your final answer here...
    </answer>

Question: {question}
```

E.2. Chain-of-thought prompting with context

```
You are a medical assistant specializing
    in surgical information. Use the
    following context to answer the
    question.
If you cannot find the answer in the
    context, say "I don't have enough
    information to answer this question."

Context:
{context_docs}

Question: {question}

Think step-by-step to reason through your
    answer and consider the relevant
    information from the context. Respond
    in the following format:
<think> Your reasoning here... </think>
<answer> The answer to the question...
    </answer>
```

E.3. ReAct Agent System Prompt

This prompt is used for the ReAct agent in the evaluation scripts:

```
You are a medical assistant specializing
    in surgical information. Your goal is
    to provide accurate and helpful
    information about surgical procedures,
    techniques, and related medical
    knowledge.

When answering questions:
1. Use the DocumentSearch tool to find
    relevant information in the surgical
    database
2. Analyze the retrieved information
    carefully
3. Provide comprehensive and accurate
    answers based on the retrieved content
4. If the information is not available in
    the search results, acknowledge the
    limitations
5. Think step-by-step to reason through
    complex questions
6. Cite specific parts of the retrieved
    documents when appropriate
7. Focus on providing factual medical
    information rather than opinions

Remember accuracy is crucial. Provide all
    reasoning and the final answer.
```

Appendix F. Prompt Design Principles

The prompts used in the Surgical Information Assistant follow several key design principles:

F.1. Step-by-Step Reasoning

All prompts encourage the language model to think step-by-step and provide detailed reasoning before arriving at an answer. This approach, often referred to as chain-of-thought prompting, has been shown to improve the accuracy and reliability of language model outputs, especially for complex tasks.

F.2. Structured Output Format

The prompts use a consistent structured output format with XML-like tags (e.g., <think>, <answer>, <sub-question>) to clearly separate different components of the response. This structured format makes it easier to parse and extract specific information from the language model's output.

F.3. Few-Shot Examples

Several prompts include few-shot examples to demonstrate the expected reasoning process and output format. These examples help guide the language model to produce responses that follow the desired pattern and meet the quality standards.

F.4. Task-Specific Instructions

Each prompt contains clear, task-specific instructions that define the role of the language model (e.g., "You are an expert at breaking complex surgical questions into simpler ones") and the specific task it needs to perform. This helps focus the model's attention on the relevant aspects of the task.

F.5. Context Awareness

The prompts are designed to make the language model aware of the context in which it operates, including the original question, retrieved information, and previous interactions. This context awareness is crucial for generating coherent and relevant responses.

Appendix G. Example responses from different experimental designs

G.1. Question 1

Why should soiled instruments be cleaned within twenty minutes of contamination?

Known answer: Soiled instruments should be cleaned within twenty minutes because stainless steel can experience corrosion from biological and chemical liquids.

Llama32 Deretsyn: The soiled instruments should be cleaned within twenty minutes of contamination to prevent the potential risks associated with them, including corrosion, injury from sharp instruments, and the spread of infection due to the presence of biological and chemical liquids. This time frame is significant because it is the time frame within which organic matter such as blood or tissue can be prevented from drying on the surface of the instrument. If not cleaned within this time frame, the consequences can include damage to the instruments, corrosion, and the potential for injury. Furthermore, there are specific guidelines that dictate the cleaning of soiled instruments within twenty minutes of contamination.

Llama32 RAG: Soiled instruments should be cleaned within twenty minutes of contamination to prevent the growth of microorganisms and the drying of organic matter, which can lead to the formation of biofilms and make the instrument more difficult to clean and disinfect.

Llama32 React: Soiled instruments should be cleaned within twenty minutes of contamination to prevent the growth and spread of microorganisms. This is because microorganisms can multiply rapidly, especially in warm, moist environments, and can cause infections if not removed promptly.

GPT4O: Soiled instruments should be cleaned within twenty minutes of contamination to prevent organic material from drying and adhering to the surface, to reduce the risk of biofilm formation that can make sterilization less effective, to minimize the risk of cross-contamination and infection, to maintain the integrity and functionality of the instruments, and to adhere to best practices in infection control.

Llama31: Soiled instruments should be cleaned within twenty minutes of contamination to prevent bacterial growth, biofilm formation, and subsequent infection transmission. This timeframe allows for effective cleaning and disinfection, reducing the risk of hospital-acquired infections and ensuring patient safety.

G.2. Question 2:

What factors can put the liver remnant at risk for further ischemic injury postoperatively after liver resection surgery?

Known answer: Hypoxia, hypotension, septicemia, or drug toxicity can put the liver remnant at risk for further ischemic injury postoperatively.

Llama32 Deretsyn: The factors that can put the liver remnant at risk for further ischemic injury postoperatively include:

1. **Hemodynamic instability**: This can affect blood flow to the liver remnant

and increase the risk of ischemic injury.
2. **Hyperdynamic circulation**: This can also affect blood flow to the liver remnant and increase the risk of ischemic injury.
3. **Septicemia**: This is a potential cause of ischemic injury in the liver remnant postoperatively.
4. **Drug toxicity**: This is another potential cause of ischemic injury in the liver remnant postoperatively.
5. **Volume expansion**: This can affect blood flow to the liver remnant and increase the risk of ischemic injury.
6. **Glycemic control**: Poor glycemic control can lead to hyperglycemia, which can contribute to ischemic injury in the liver remnant.
7. **Liver regeneration**: While liver regeneration is a natural process, it can also lead to changes in blood flow to the liver remnant, increasing the risk of ischemic injury.
8. **Ascites**: This can put pressure on the liver remnant and impair venous return to the heart, increasing the risk of ischemic injury.
9. **Bleeding**: This can lead to hemodynamic instability and affect blood flow to the liver remnant, increasing the risk of ischemic injury.
10. **Portal hypertension**: This can affect blood flow to the liver remnant and increase the risk of ischemic injury.
11. **Cardiac function**: Poor cardiac function can lead to hypotension and affect blood flow to the liver remnant, increasing the risk of ischemic injury.
12. **Coagulation status**: Abnormal coagulation status can lead to bleeding and affect blood flow to the liver remnant, increasing the risk of ischemic injury.
13. **Warm ischemia time**: Prolonged warm ischemia time can increase the risk of ischemic injury in the liver remnant.
14. **Liver function**: Poor liver function can affect blood flow to the liver remnant and increase the risk of ischemic injury.
15. **Postoperative factors**: Hypoxia, hypotension, septicemia, drug toxicity, hyperdynamic circulation, and hyperglycemia can all contribute to ischemic injury in the liver remnant postoperatively.
16. **Surgical procedure**: The surgical procedure itself can increase the risk of hypoxia and hypotension, prolong warm ischemia time, and cause hemodynamic instability, all of which can increase the risk of ischemic injury in the liver remnant.

These factors can all contribute to ischemic injury in the liver remnant postoperatively, either by affecting blood flow to the liver remnant or by increasing the risk of hypoxia, hypotension, septicemia, and drug toxicity.

Llama32 RAG: Hypoxia, hypotension, septicemia or drug toxicity, and hyperdynamic circulation due to cirrhosis or impaired liver function can put the liver remnant at risk for further ischemic injury postoperatively.

Llama32 React: The following factors can put the liver remnant at risk for further ischemic injury postoperatively: 1. Warm ischemia time, 2. Fatty or cirrhotic livers, 3. Large resections, 4. Prolonged surgical time, 5. Blood loss and cardiovascular instability, 6. Non-steroidal anti-inflammatory drugs (NSAIDs), 7. Paracetamol (acetaminophen) use, 8. Hypoglycemia, 9. Hypothermia, 10. Poor patient selection.

GPT4O: Factors that can put the liver remnant at risk for further ischemic injury postoperatively include vascular and perfusion issues (e.g., portal vein thrombosis, hepatic artery injury, small-for-size syndrome), systemic and local hypoxia, inflammatory and oxidative stress related to ischemia-reperfusion injury, pre-existing conditions that impair liver function (e.g., cirrhosis, steatosis, diabetes), poor postoperative hemodynamic management (e.g., hypotension, inadequate resuscitation), and a small or insufficiently functional liver remnant that cannot meet metabolic demands.

Llama31: The liver remnant is at risk for further ischemic injury postoperatively due to a combination of factors, including:

- Hemodynamic factors (e.g., changes in blood flow, blood pressure, or cardiac output)
- Surgical factors (e.g., extent of liver resection, presence of major blood vessels in the area of resection)
- Patient-related factors (e.g., overall health, age, presence of comorbidities)
- Specific conditions (e.g., small liver remnant, portal hypertension, biliary obstruction, dehydration or hypovolemia)
- Inadequate intraoperative and postoperative care (e.g., fluid management, blood pressure management, monitoring and management of complications)